International Journal of Research in Engineering, Science and Management
Volume 5, Issue 11, November 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Analysis of Various Machine Learning Algorithms on Wind Power Forecasting with Introduction of Varying Rates of Missing Data

Rakshith Dasenahalli Lingaraju*

Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

***Abstract*: Machine learning (ML) is a process which reads through data and finds statistical relationships and structure within the data which can be used to make predictions. Various industry sectors have implemented ML and are reaping great benefits. However, the energy sector has been slow in its adoption and has yet to see its full potential. This paper aims to address how Machine Learning can greatly benefit the energy sector in optimizing and improving accuracy of wind power forecasting. This paper compares performance of two algorithms: Support Vector Regression (SVR) [1] and K-Nearest Neighbor (KNN) [2] in wind power forecasting and will be evaluated based on the Mean Square Error (MSE). The true predictive capability of the trained model will be analyzed when data points at random are removed to test their effect on MSE of the algorithms.**

***Keywords*: Machine Learning, K Nearest Neighbor, Support Vector Regression, Mean Square Error, Wind power forecasting, Randomized missing data.**

1. Introduction

Wind power is the use of air flow through wind turbines to generate power. Wind is the second most renewable energy [3] used in the generation of electricity after Hydropower. It is a clean source of energy, pollution free and produces no greenhouse gasses. Wind farms are a collection of wind turbines connected to the electric grid. In 2020 Wind power accounted for just 6% of total power generated in the world. However, according to the International Energy Agency, it is growing at a rate of about 17% annually [4]. Wind farms are also growing rapidly all around the world and are set to represent a significant part of future energy sources. Cost of wind turbines are also coming down thus enabling their further growth. Biggest hurdles in its larger implementation is its fluctuation and seasonality. This affects the overall efficiency of the wind turbine and makes it unreliable as a stable source of energy.

This can be overcome by using Machine learning algorithms to predict its future variability within a small margin of error. Algorithms are also used to sift through huge amounts of seemingly cluttered data to identify intrinsic signals hidden within it. This helps in understanding the data better to make accurate predictions.

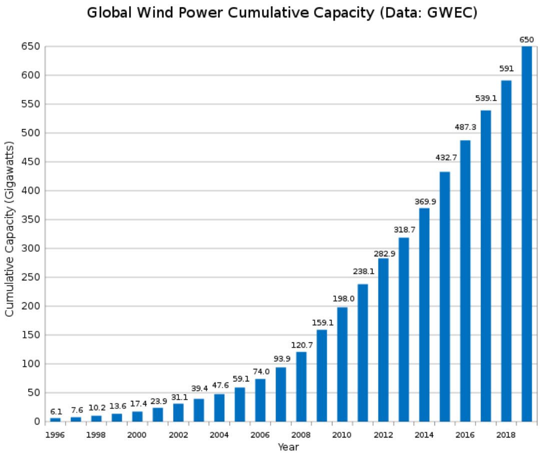


Fig. 1. Global wind power cumulative capacity data [5]

Regression is a method or a technique in which the relationship between independent features and dependent features are established. It is a commonly used type of machine learning, to predict outcomes which are continuous. In this paper we cover the use of regression algorithms to predict wind power. Although machine Learning is not widely used in the Power industry, there is a tremendous opportunity that ML offers which the industry can leverage. Other industries that do leverage the value of ML are decades ahead, yet the energy sector has failed to catch on.

Machine learning algorithms generally have multiple steps, namely, training, testing and validation. To achieve this, the entire data is split into these parts respectively. During training, the ML model is trained to identify a best fit curve for the training data and then tested to ensure the algorithm can perform well even on unseen data.

2. Data Collection

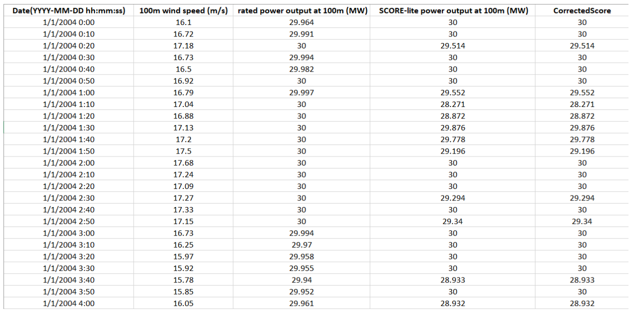
National Renewable Energy Laboratory (NREL) [6] is the only federal laboratory in the US specializing in renewable energy and efficiency. As part of their Research & Development, they extensively collect huge amounts of Data in Solar, Wind, and other renewable sources. This is done by the

*Corresponding author: rakshithdl@gmail.com

Renewable Resource Data Center (RReDC) [7]. They also detail their resource information through maps, reports, tools and process of data control.

In Wind, NREL has collected data spanning over several years as far back 2004 from multiple Wind farm locations like Reno, Tehachapi etc. They represent big data, and the data is collected every 10 seconds and stored in a .xml format. The focus in this paper was mainly on two parts of the data: Wind speed and power. Since they are correlated, wind speed will act as the x-axis data while Power will act as Y-axis. Also since the data is a time series data, time is also a variable considered. The data is collected at an altitude of 100m.

Once the data is collected pre-processing steps have to be performed to ensure that data is not corrupted, and a standard format is maintained. Any errors at this stage will not let the algorithm perform efficiently. If any errors occur, they are dealt with, either by removing them completely, or changing them into a format that is more comfortable to work with.



Date(YYYY-MM-DD hh:mm:ss)	100m wind speed (m/s)	rated power output at 100m (MW)	SCORE-lite power output at 100m (MW)	CorrectedScore
1/1/2004 0:00	16.1	29.964	30	30
1/1/2004 0:10	16.72	29.991	30	30
1/1/2004 0:20	17.18	30	29.514	29.514
1/1/2004 0:30	16.73	29.994	30	30
1/1/2004 0:40	16.5	29.982	30	30
1/1/2004 0:50	16.92	30	30	30
1/1/2004 1:00	16.79	29.997	29.552	29.552
1/1/2004 1:10	17.04	30	28.271	28.271
1/1/2004 1:20	16.88	30	28.872	28.872
1/1/2004 1:30	17.13	30	29.876	29.876
1/1/2004 1:40	17.2	30	29.778	29.778
1/1/2004 1:50	17.5	30	29.196	29.196
1/1/2004 2:00	17.68	30	30	30
1/1/2004 2:10	17.24	30	30	30
1/1/2004 2:20	17.09	30	30	30
1/1/2004 2:30	17.27	30	29.294	29.294
1/1/2004 2:40	17.33	30	30	30
1/1/2004 2:50	17.15	30	29.34	29.34
1/1/2004 3:00	16.73	29.994	30	30
1/1/2004 3:10	16.25	29.97	30	30
1/1/2004 3:20	15.97	29.958	30	30
1/1/2004 3:30	15.92	29.955	30	30
1/1/2004 3:40	15.78	29.94	28.933	28.933
1/1/2004 3:50	15.85	29.952	30	30
1/1/2004 4:00	16.05	29.961	28.932	28.932

Fig. 2. NREL data sample for wind power [8]

Since the data is collected for research purposes, NREL has ensured that there is no corrupted or missing data. This is quite good from a research point of view, as experiments can be performed to get the best result from this ideal representation of data. This helps in setting up a baseline, which can be used as a reference to compare against when the model is tested on real world data.

3. Wind Farms

This paper addresses data collected from 4 Wind farms:

- Tehachapi
- Cheyenne
- Palm Springs
- Reno

1) Tehachapi (California)

One of the initial large scale wind farms installed in the US in the 1980s. Tehachapi is a city in California at 1210m above sea level. They were built to transport power to different parts of California and underwent several upgrades in both 2008 and 2012.

2) Cheyenne (Wyoming)

Cheyenne is the Capital and one of the most populous cities of the State of Wyoming in the United States of America. It's at an elevation of 1848m with moderate to high wind speeds.

3) Palm Springs (California)

It is a desert resort city in California, United states of America at a low elevation of 146m with maximum hub height at 160 ft. It has in operation 3000+ units.

4) Reno (Nevada)

It's a City in Nevada close to Lake Tahoe with an elevation of 1373m supplying power to many parts of Nevada. It supports Wind power generation as it has abundant wind energy and excellent wind incentives and rebates.

4. Algorithms

Algorithm is a specification of how to solve a class of problems. Machine Learning also has various algorithms in its repertoire which deals with the problem in different ways. Hence, it's important to decide the type of algorithm to use in a given situation, which depends entirely on the type of data and the results which are expected.

This paper compares the performance of two algorithms on predicting wind power: Support Vector Regression (SVR) and K Nearest Neighbor (KNN).

1) Support Vector Regression (SVR)

It is a supervised learning method, which needs trained labels in data to learn from. This is exactly the type of data which was collected from NREL, where the Power values in the data represent the labels. In simplistic terms it is a linear classifier where it divides data into different classes based on how the data is clustered.

As seen in Figure 3, there are 2 classes, which can be classified by drawing a linear line which acts as a Separator. Lines can be drawn in any fashion, but the best separator is the one which has maximum separation between the closest points on either side of the Separator.

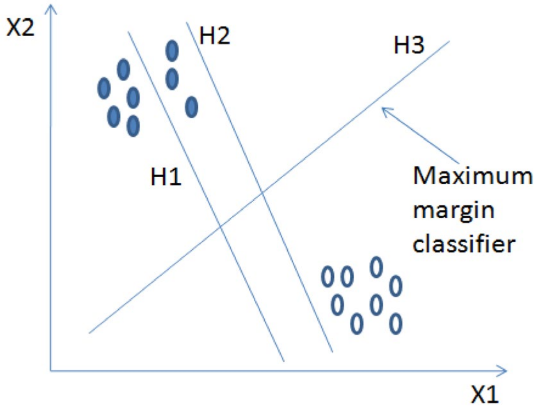


Fig. 3. Representation of Support Vector Regression (SVR) [9]

In Figure 3, we see 3 lines H1, H2 and H3 that act as separators. H1 goes through the cluster, hence it's not a good separator and is rejected. Both H2 and H3 correctly split data into separate classes, but compared to H2, H3 is in such a way that there is maximum separation between the closest points on either side of H3. Hence H3 is chosen as the separator. It is also clear that the choice of separator depends on the closest point to the separator on either side. This creates issues for data that is largely spread out, as a single data point can decide how the classes are split.

2) K Nearest Neighbor Regression (KNN)

It's based on proximity voting, where the class/label of each data point is decided by choosing "K" neighbor points, and then seeing which class most of those neighbor points belong to. Based on that, the selected point takes that class. "K" is typically chosen to be an odd number to avoid a tie in voting, especially in binary classification.

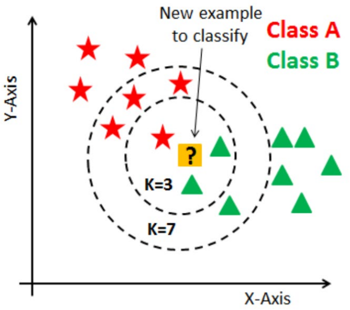


Fig. 4. Representation of K Nearest Neighbor (KNN) [10]

In Figure 4, "New example to classify" is the point of interest, whose class needs to be determined. If K is chosen to be 3, there exists 1 red star and 2 green triangle data points as its neighbor. By vote of majority, the chosen class is Green.

If K=7, observation can be made that there are 4 red stars and 3 green triangles as its neighbor which changes the assigned class to Red. Since the value of K can strongly affect the outcome, selecting an optimum value of K becomes vital.

5. Measure of Error

To determine the performance of the algorithm, its error in prediction compared to the actual value needs to be measured. In this paper, mean squared error (MSE) [11] was selected, which measures the average of the square of the error, error being the difference between the predicted wind power to the actual power generated.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(predicted - actual)^2$$

where n is the number of predictions.

Fig. 5. Mathematical formula of mean squared error (MSE) [11]

6. Comparing the Performance of Algorithms

Analysis of algorithms is performed between SVR and KNN, and a better performing algorithm is one with the lowest Mean Squared Error (MSE). Ideally, error should be zero, but practically there will always exist some error, and the aim is to find an algorithm with the least error.

The analysis is done on all the 4 wind farms and as seen in Figure 6. For the time duration of 12 months, KNN algorithm has lower MSE for Tehachapi, Palm springs and Reno, while SVR has the lower MSE for Cheyenne. The results vary if a different duration of time is chosen.

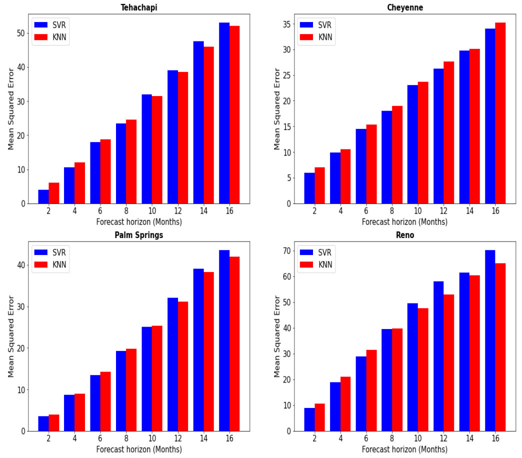


Fig. 6. MSE plot for SVR and KNN for different wind farms

7. Forecasting Wind Power

Research conducted in this paper compares the performance of the algorithms to forecast the Wind power with very low MSE. Lower MSE helps in increasing the reliability of the forecast and to better prepare the power grid to accommodate the variability of the wind power generation.

In figure 8, 9, 10 and 11, plots of predicted values are overlaid with the actual values. All 4 wind farms have different profiles, which is to be expected as they are in different parts of the United States of America. Observations can also be made that the overall prediction generally matches with the actual values, except for regions where sudden spikes occur. This confirms that the model was not overtrained and is able to effectively handle new unseen data.

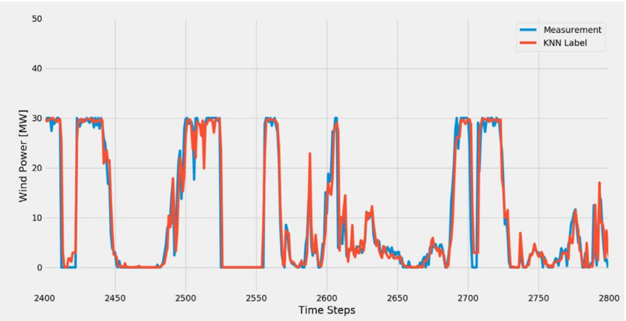


Fig. 8. Wind forecasting for Tehachapi wind farm

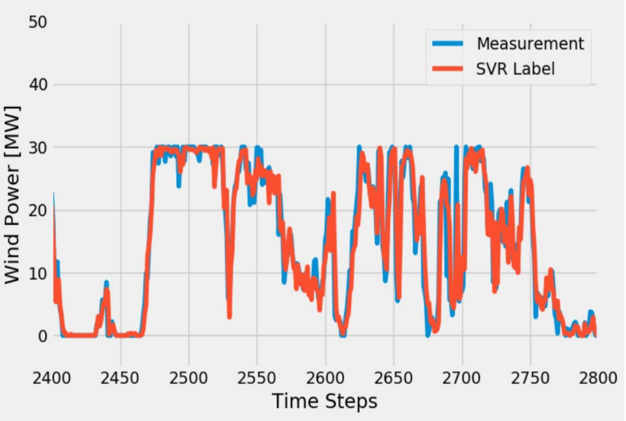


Fig. 9. Wind forecasting for Cheyenne wind farm

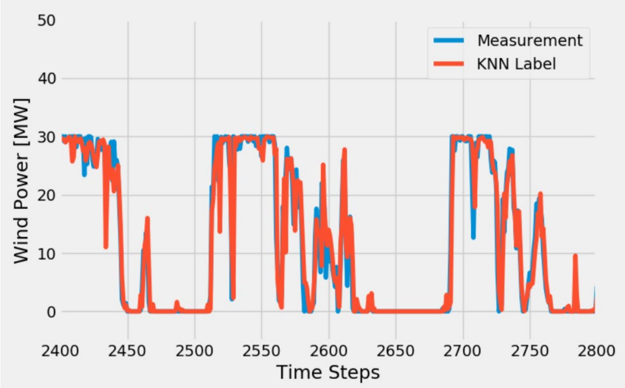


Fig. 10. Wind forecasting for Palm Springs wind farm

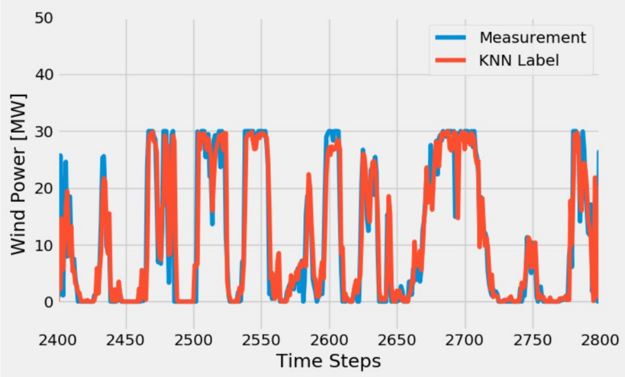


Fig. 11. Wind forecasting for Reno wind farm

The wind power data obtained from NREL was filtered for 2 years, which was split into 2 parts: training and testing data. The first-year data was used to train the model and the second-year data for testing. During training, both input variable (named x) and outcome variable (named y) i.e., Wind speed and Power generated are used and the algorithm/model parses through this data to find a best fit curve. Several iterations and hyperparameters of the model are run to identify the model with the least error.

Once the model is trained, the unseen test data is scored by the model to predict the wind power generated. The predicted values are then compared with actual values to compute the MSE (Mean Squared Error). This is performed individually for all the 4 wind farms.

8. Missing Data

The data collected by NREL from these wind farms are ideal since the data was for research purposes. Hence great care was taken to ensure that data was collected without any missing data or noise. This establishes a baseline of what to expect when data is clean and appears as expected.

However, only testing on such ideal data is not useful, since in the real world, data will likely be either missing for some duration, or have random noise introduced etc. Results obtained from such ideal data are not reliable since the algorithms are not trained to handle data with errors or missing data points. In this paper, model outcomes for ideal data and missing data are compared to understand the effects of missing data on the performance of ML models and their overall results.

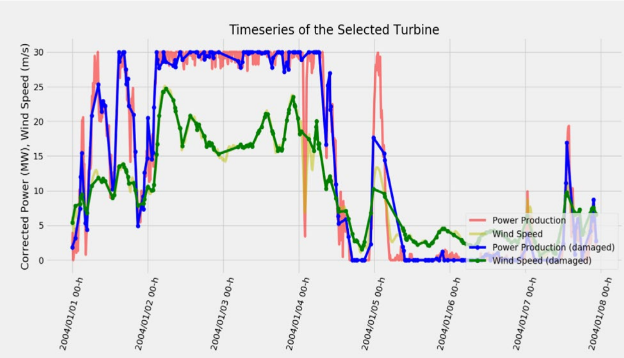


Fig. 12. Missing data representation

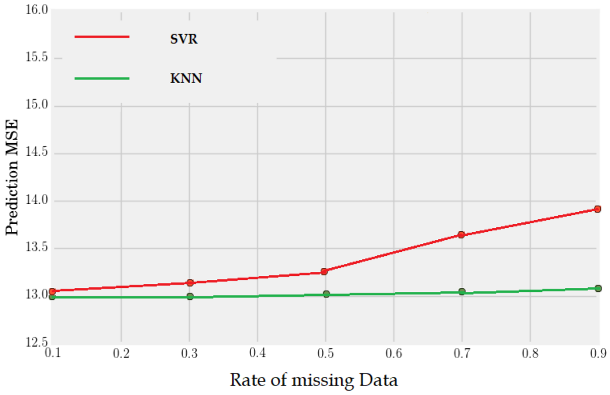


Fig. 13. Effect of rate of missing data on MSE

In figure 13, it is clearly seen that MSE for KNN remains nearly constant for any rate of missing data, while for SVR it increases drastically for rate above 0.5. This can be clearly explained by how each of these algorithm's work.

In the case of SVR, the separator depends on the nearest point on either side of it, and at a higher rate of missing data, it is highly likely that the nearest point itself was missing, thus affecting the separator selection and in turn increasing MSE. On the other hand, for KNN, if the proper "K" value is chosen, even with few neighboring data points missing, considering majority voting, results will likely not change by much, and hence why their MSE remains fairly constant.

9. Visualization

Visualization is a powerful tool which if used properly can help to better understand the data. It becomes much easier to understand the context of the model behavior visualized than to see the exact same data as numbers.

Figure 14 is a representation of different turbines from the Tehachapi wind farm over a period of time. Such data is called time series data and its representation is called time series representation. It makes it easy to see the peaks and troughs of the power generated throughout the timeline and also variations among different turbines within the same wind farm. This might be due to a multitude of factors, including hub height difference, location in the wind farm etc.

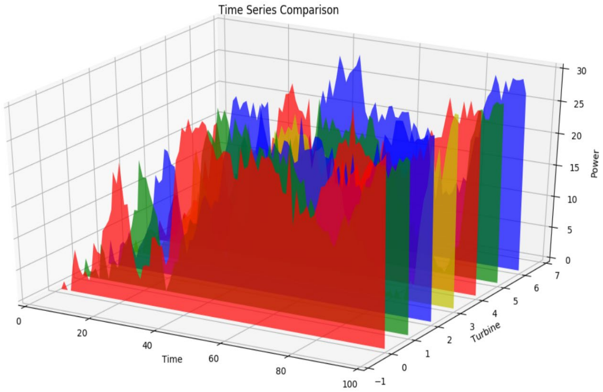


Fig. 9. Visual representation of power generation of wind turbines

10. Software Used

Multiple software packages were relied upon for this research.

1) Python IDLE

It's the Software in which the code was executed to generate results. The programming language that was used was Python with version 2.7.

2) Brunel

It's a visualization package which is supported by Python and is a powerful tool in representing machine learning algorithm results.

3) Matplotlib

Another visualization tool available as a package within Matlab, used in representing graphical plots.

4) Sklearn

It's a Machine Learning package supported by python. The advantage of such a code package, is that the algorithm doesn't need to be built from scratch and helps in prototyping quickly.

11. Advantages of Utilizing ML Models

- Higher accuracy than traditional methods at forecasting.
- Robust to missing data.
- Helps to optimize and create a more efficient system.
- Able to deal with big and complex data easily.

12. Future Scope

- Utilize the same methodology in other power sectors like Solar, Thermal etc.
- Use prediction values of wind power generation to optimize and create a stable power grid.
- Analyze the results with additional algorithms to improve results.

13. Conclusion

Machine Learning algorithms have shown tremendous results in accurately forecasting wind power. It has also shown how effectively it can accurately predict wind power even with unforeseen and messy data. This will help the energy sector achieve greater efficiency and optimization and leads to creating a more stable power grid. There are still many fields in the energy sector where machine learning can help, and with more research in the coming decades it is sure to see some groundbreaking improvements.

Acknowledgement

Special thanks to Professor Mohammed Beshir, Department of Electrical and Computer Engineering Practice, USC for his guidance and support during this research work.

References

[1] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

[2] Padraig Cunningham, Sarah Jane Delany. k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). https://arxiv.org/abs/2004.04523

[3] https://www.c2es.org/content/renewable-energy/

[4] https://www.iea.org/reports/wind-electricity

[5] https://upload.wikimedia.org/wikipedia/commons/1/1d/Global_Wind_Power_Cumulative_Capacity.svg

[6] https://en.wikipedia.org/wiki/National_Renewable_Energy_Laboratory

[7] https://www.nrel.gov/rredc/

[8] http://vegas.informatik.uni-oldenburg.de/data/nrel/2004/

[9] https://www.mdpi.com/water/water-13-01172/article_deploy/html/images/water-13-01172-g004.png

[10] https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8

[11] (2011). Mean Squared Error. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.